Estudio Comparativo y Nuevas Propuestas de Técnicas de Parametrización de la Señal de Voz para el Reconocimiento del Habla

Jaume Clot, Javier Hernando, Climent Nadeu y Francesc Vallverdú
Departament de Teoria del Senyal i Comunicacions
Universitat Politècnica de Catalunya
Ap. 30002, 08071 Barcelona
Tel: (93) 401 64 40 Fax: (93) 401 64 47 E-mail:javier@tsc.upc.es

Abstract.- A correct choice of voice signal modeling method is essential to obtain good results in automatic speech recogniton. In this paper, a comparative study betwen two speech signal models, Linear Prediction Coeficients and mel-cepstrum filter bank, is presented. On the other hand, a new parameterization method, obtained as a hybrid from previous models (LPC-mel-cepstrum), is presented too.

1. Introducción

El reconocimiento del habla requiere como primer paso la conversión de la señal de voz en una secuencia temporal de espectros correspondientes a segmentos consecutivos de señal, que responden a un modelo matemático caracterizado por un número limitado de parámetros (parametrización). Posteriormente, dicha secuencia de espectros es confrontada con otras en base a técnicas de reconocimiento de formas a fin de clasificarla.

Sin embargo, con las técnicas actuales, sólo en el caso de palabras aisladas, cuando el vocabulario es reducido y en situaciones acústico-fonéticas (variabilidad fonética, tipo de locutores, ruido ambiental, ...) poco dificultosas, existen soluciones satisfactorias. Una mejora de los métodos de extracción de características de la señal de voz es necesaria para aumentar las prestaciones de los sistemas actuales de reconocimiento.

En la mayoría de aplicaciones de procesado del habla, la señal de voz suele modelarse como la respuesta de un filtro todopolos variante en el tiempo a una señal de excitación cuyo espectro puede ser plano (segmento sordo) o un tren de líneas espectrales situadas a frecuencias múltiplos de la frecuencia fundamental o tono de voz (segmento sonoro). Por tanto, la envolvente del espectro de la señal de voz, que contiene la mayor parte de la información del mensaje oral, se corresponde con la respuesta frecuencial del filtro.

Los métodos de parametrización de la señal de voz usados en reconocimiento tienen por objeto estimar dicha envolvente espectral. En la resolución de este problema, pueden distinguirse claramente dos enfoques. En primer lugar, los métodos que tratan de determinar los coeficientes del filtro del modelo anterior, usualmente por *predicción lineal (LPC)*. En segundo lugar, los métodos que obtienen una estimación de la envolvente espectral eliminando el contenido de la señal de excitación a dicho filtro mediante un simple proceso de integración en el tiempo *(técnica de banco de filtros)* o en frecuencia *(parametrización mel-cepstrum)*.

El propósito de este trabajo es doble: *1)* realizar un estudio comparativo de los dos métodos de parametrización más usados en los sistemas actuales de reconocimiento del habla, *LPC* y *mel-cepstrum*; *2)* proponer un método híbrido consistente en la aplicación de la técnica *mel-cepstrum* sobre la envolvente espectral obtenida mediante *predicción lineal*.

La comunicación está organizada del siguiente modo. En el apartado 2 se revisan los métodos de parametrización que son objeto del estudio comparativo y se presenta el nuevo método de modelado de la señal de voz. En el apartado 3 se describen las pruebas experimentales y se muestran los resultados de reconocimiento obtenidos. Por último, se exponen algunas conclusiones.

2. Métodos de Parametrización

2.1 Predicción Lineal

La técnica de *predicción lineal* es ampliamente utilizada en tratamiento del habla y equivale a un modelado autoregresivo de la señal de voz. Usualmente se estiman los coeficientes del filtro todopolos del modelo de predicción de voz aplicando el algoritmo de Levinson-Durbin [1] sobre la secuencia de autocorrelación de la señal. A partir de estos coeficientes, se calculan también de forma recursiva, los coeficientes *cepstrum*.

2.2 Mel-Cepstrum

En la técnica *mel-cepstrum,* propuesta por S.B.Davis y P.Mermelstein [2], se intenta eliminar la contribución al filtro del modelo mediante una integración en el dominio frecuencial. Para ello, en primer lugar se calcula el periodograma del tramo de señal, módulo al cuadrado de la DFT del tramo. Posteriormente, se aplica sobre este espectro un banco de filtros triangulares paso-banda distribuidos uniformemente en escala *mel* (1) sobre el rango de frecuencias deseado (ver figura 1).

$$mel\ frequency = 2595 \log 10(1+f/700) \qquad (1)$$

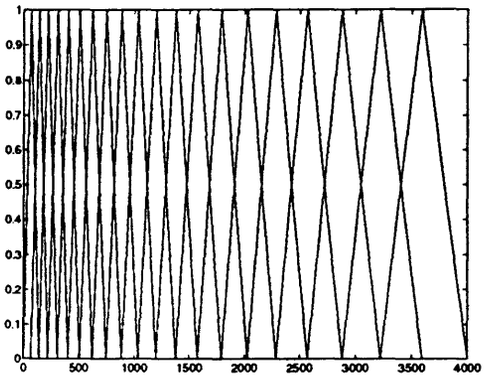

Figura 1. Banco de filtros en escala *mel*

Los valores, en escala logarítmica, resultantes de integrar las muestras ponderadas del periodograma en cada banda de frecuencias, se utilizan para calcular los coeficientes *mel-cepstrum* a través de la transformada coseno inversa. Como resultado de este proceso se obtienen unos coeficientes que tienen en cuenta la sensibilidad logarítmica en intensidad y frecuencia del oído y su resolución en bandas críticas. Precisamente cada filtro *mel* corresponde con una banda crítica, aproximadamente 20 en un rango de frecuencias de 4 KHz.

2.3 Método híbrido LPC-mel-cepstrum

El método híbrido de parametrización propuesto es el siguiente. En primer lugar obtenemos los coeficientes del filtro del modelo de producción de voz a partir del algoritmo de Levinson-Durbin de *predicción lineal* y calculamos el módulo al cuadrado de la respuesta frecuencial del filtro. Esta función espectral, denominada espectro *LPC,* es una estimación de la envolvente del espectro de la señal de voz, ya que se supone que la envolvente espectral de la señal de excitación al filtro es plana. Posteriormente, sobre el espectro *LPC* aplicamos la técnica *mel-cepstrum*. En este modelado híbrido se combinan, pues, las dos aproximaciones

al problema de la estimación de la envolvente espectral de la señal de voz: determinación de los coeficientes del filtro del modelo de producción de voz y eliminación del contenido de la señal de excitación a dicho filtro mediante un proceso de integración, en este caso en frecuencia.

3. Resultados experimentales

3.1 Base de datos y sistema de reconocimiento

Las pruebas se han realizado sobre la base de datos TI [3] que consta de un gran número de secuencias de dígitos en inglés pronunciadas en ausencia de ruido. Esta base de datos se proporciona muestreada a 20 KHz. Para este trabajo se interpoló por 2 y se diezmó por 5 a fin de obtener una señal muestreada a 8 KHz. Se utilizaron sólo a los locutores adultos de la base (112 de entrenamiento y 113 de test) y de ellos sólo se usaron los dígitos aislados, 22 por locutor.

El sistema de reconocimiento usado es de reciente aparición en el mercado (HTK v1.5) [4] y está basado en Modelos de Markov ocultos de densidad continua [5]. En la etapa de parametrización se aplicó una ventana de *Hamming* sobre la señal de 30 ms de duración a intervalos de 10 ms. Se realizó preénfasis con factor 1.

Cada palabra se caracterizó por un Modelo de Markov de 10 estados de izquierda a derecha, sin saltos, con una matriz de covarianza diagonal y una mezcla por estado. Se utilizó un modelo para el silencio del mismo tipo pero con 5 estados.

El entrenamiento de los modelos se realizó en dos etapas: una etapa de inicialización mediante el algoritmo *Segmental K-means*, con segmentación manual previa, y una etapa de reestimación mediante el algoritmo de *Baum-Welch*. En la fase de test se utilizó un algoritmo de *Viterbi* clásico.

3.2 Resultados de reconocimiento

A continuación se muestran los resultados obtenidos para los diferentes métodos de parametrización y variando el orden p del modelo y el número N de coeficientes *cepstrales*. En el caso de la parametrización *LPC* y para el método híbrido, el orden se refiere al número de polos del filtro del modelo y en el caso de *mel-cepstrum*, al número de filtros empleado para modelar los coeficientes. En la técnica de *predicción lineal* el orden de predicción ha de ser mayor o igual al número de coeficientes *cepstrales* si se quiere que caractericen completamente al modelo (ver tabla 1). En cuanto a la técnica *mel-cepstrum* el número de filtros ha de ser mayor que el número de coeficientes *cepstrum* debido a las características de la transformada coseno (ver tabla 2). En el método híbrido propuesto, existen dos órdenes: el orden de predicción y el número de filtros en la escala *mel*. Tanto para los órdenes de análisis como para el número de coeficientes se ha optado por considerar únicamente números pares para disminuir el número de pruebas. Los resultados corresponden al número de reconocimientos erróneos de un total de 2468.

En cuanto a la parametrización LPC, se observa en la tabla 1 que los valores de p y N más usados (orden de análisis 10 y 12 coeficientes *cepstrales*) no proporcionan en esta aplicación, los mejores resultados. El menor número de errores se obtiene utilizando un orden *LPC* entre 14 y 18, y un número de coeficientes entre 18 y 22. Cabe destacar que el orden que proporciona mejores resultados, independientemente del número de coeficientes, es 18 y el número de coeficientes *cepstrum* que consigue también los mejores resultados, independientemente del orden *LPC*, es 20, siendo los valores óptimos para nuestro caso orden 18 y 26 coeficientes (69 errores).

En el caso de la técnica *mel-cepstrum*, observamos que los mejores resultados se obtienen usando 8 coeficientes *cepstrum* siendo el número de filtros no demasiado crítico en esta aplicación. El número de filtros más utilizado, 20, ha resultado ser uno de los que ofrece menores tasas de error, y observamos mínimos en reconocimientos erróneos para 18 y 26 filtros con 8 coeficientes (60 errores).

L P C															
	número de coeficientes cepstrum														
		8	10	12	14	16	18	20	22	24	26	28	30	32	34
orden	8	104	98	98	97	76	70	75	89	102	100	106	114	120	127
	10	.	112	86	85	75	78	73	86	86	93	101	109	121	138
	12	.	.	103	95	76	83	88	92	89	83	91	100	113	126
	14	.	.	.	87	78	75	72	72	76	74	84	86	87	105
	16	.	.	.	.	89	78	72	80	85	84	80	82	80	83
	18	.	.	.	.	.	74	71	75	75	69	70	74	76	81
	20	.	.	.	.	.	.	73	81	81	79	72	78	78	81
	22	.	.	.	.	.	.	.	78	84	82	79	77	86	88
	24	.	.	.	.	.	.	.	.	80	80	80	82	85	84
	26	.	.	.	.	.	.	.	.	.	80	85	97	97	96
	28	.	.	.	.	.	.	.	.	.	.	83	92	96	106
	30	.	.	.	.	.	.	.	.	.	.	.	88	91	102
	32	.	.	.	.	.	.	.	.	.	.	.	.	95	108
	34	.	.	.	.	.	.	.	.	.	.	.	.	.	96

Tabla 1. Número de errores. Parametrización *LPC*.

Mel-cepstrum						
	número de coeficientes cepstrum					
		6	8	10	12	14
orden	14	82	83	90	96	.
	16	73	76	88	91	93
	18	68	60	73	89	88
	20	74	63	71	68	102
	22	75	63	76	68	86
	24	74	62	82	74	92
	26	76	60	75	74	89
	28	72	65	80	78	87
	30	88	73	71	77	97
	32	84	68	69	76	79
	34	87	68	77	77	91
	36	83	68	75	74	84

Tabla 2. Número de errores. Parametrización *Mel-cepstrum*.

LPC-mel-cepstrum									
		20 filtros			30 filtros			50 filtros	
	número de coeficientes cepstrum								
		6	8	10	6	8	10	6	8	10
orden LPC	8	82	88	92	90	91	98	86	88	103
	10	67	75	83	72	74	83	74	79	84
	12	71	77	86	73	76	88	74	79	89
	14	69	74	86	72	82	80	80	78	78
	16	69	72	77	75	82	80	86	76	73
	18	70	75	77	76	72	79	87	71	78
	20	80	79	80	84	75	82	89	69	75
	22	81	72	84	82	77	79	88	73	73

Tabla 3. Número de errores. Parametrización *LPC-mel-cepstrum*.

En cuanto al modelo híbrido, puede observarse que se necesitan menos coeficientes para representar el espectro, siendo 6 el número que ofrece mejores resultados en nuestras pruebas (67 errores), con 20 filtros *mel* y orden *LPC* 10. Esta reducción del número de coeficientes puede deberse al hecho de que se ha aplanado doblemente el espectro en dos etapas sucesivas, *predicción lineal* e integración en frecuencia.

4.Conclusiones

En el reconocimiento de dígitos aislados, utilizando la base de datos TI, la técnica de parametrización que ha ofrecido mejores tasas de reconocimiento es la *mel-cepstrum*, que consigue un mínimo de 60 errores de reconocimiento (2.43% de tasa de error) en el caso de 18 filtros y 8 coeficientes, mientras que el método *LPC* nos ofrece como mínimo 69 errores de reconocimiento (2.80% de tasa de error) con un orden de parametrización de 18 y 26 coeficientes. El método híbrido *LPC-mel-cepstrum* desarrollado en este trabajo, proporciona un resultado intermedio con un mínimo en 67 errores (2.71% de tasa de error). Sin embargo, este resultado lo consigue con un menor número de coeficientes, 6, lo cual facilita las tareas computacionales de entrenamiento de modelos y reconocimiento. Actualmente seguimos investigando en nuevos métodos de hibridación.

Referencias

[1] John Makhoul, "Linear Prediction: A Tutorial Review" Proceedings of the IEEE, vol. 63, nº 4, Abril 1975, pp. 561-580.

[2] S.B.Davis, P.Mermelstein, "Comparision of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences" IEEE Trans. ASSP, vol. 28, 1980, pp. 357-366

[3] R.G.Leonard, "A database for speaker-independent digit recognition" Proc. ICASSP´84, Marzo 1984, pp.42.11.1-4.

[4] Cambridge University Engineering Departament Speech Group and Entropic Research Laboratories Inc."HTK - Hidden Markov Model Toolkit v1.5" Deciembre 1993.

[5] Lawrence R. Rabiner, "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition" Proceedings of the IEEE, vol. 77, nº 2, Febrero 1989, pp. 257-285